Using character n-grams to classify native language in a non-native English corpus of transcribed speech

Using character n-grams to classify native language in a non-native English corpus of transcribed speech

A central issue in second language acquisition research is the degree to which first language (L1) has an effect on the learning of a second language (L2). Recently, corpus-based computational methods have been employed to mine the language of L2 learners in order to detect such effects over large datasets. An important advancement in this area was the claim that a speaker's L1 phonology may have an effect on word choice in L2 [1], a hypothesis formulated from patterns observed in a corpus of writing of non-native English speakers.

This paper extends the analysis from L2 writing (as in [1]) to L2 naturalistic English speech, in which phonological effects might be stronger than in writing. The speech database comes from the Wildcat corpus of native- and foreign-accented speech [2], part of which was gathered using a referential communication task between dyads. Confirming the results in [1], a k-nearest neighbors classifier operating on character n-grams performs well above chance in predicting the native language of speakers (English, Korean, or Chinese).

The paper explores the relative contributions of specific word choices and phonological constraints to the character n-gram patterns. The classifier maintains high performance when highly frequent n-grams and words are removed, a strategy to control for function word statistics as a reflex of L1 background. Initial observations suggest that the effects of content words involve both specific lexical substitutions as well as systematic avoidance of words with problematic sequences. These effects are further examined through additional statistical classifiers. An advantage of this method for acquisition research is the non-reliance on linguistic errors to identify L1 effects on L2; rather, the approach allows for the analysis of more gradient effects of dispreference and avoidance. Overall, our results demonstrate how text-based corpus methods may be usefully applied to transcripts of naturalistic speech.

[1] Tsur, Oren and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 6-16, Prague, Czech Republic, June 2007.

[2] Van Engen, Kristin, Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. In press. The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*.